HYPOTHESIS TESTING

Zack Kertcher

Statistics for Management
Fall 2016

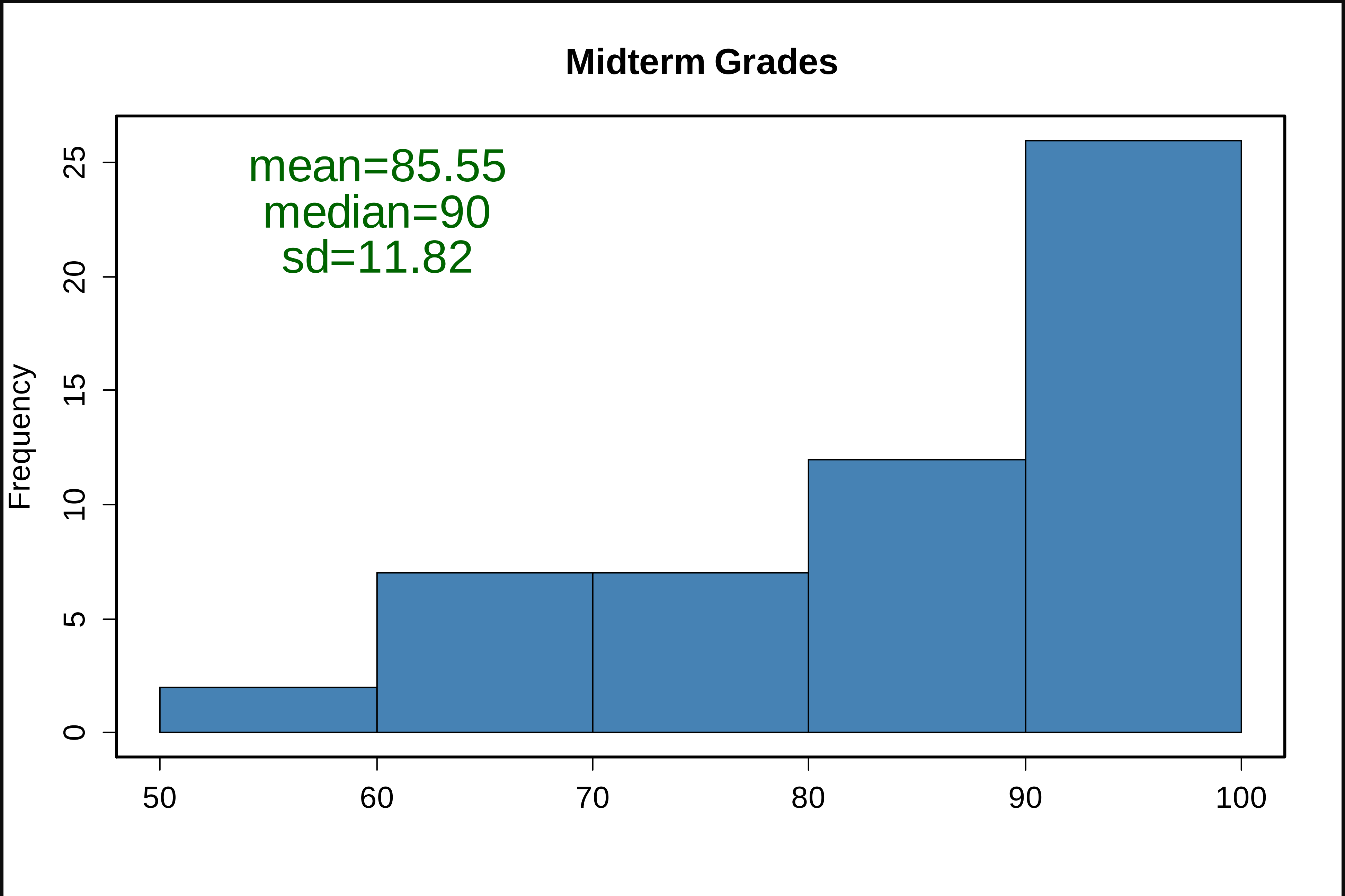
Plan for today

1. Distributions

2. Sampling and confidence interval

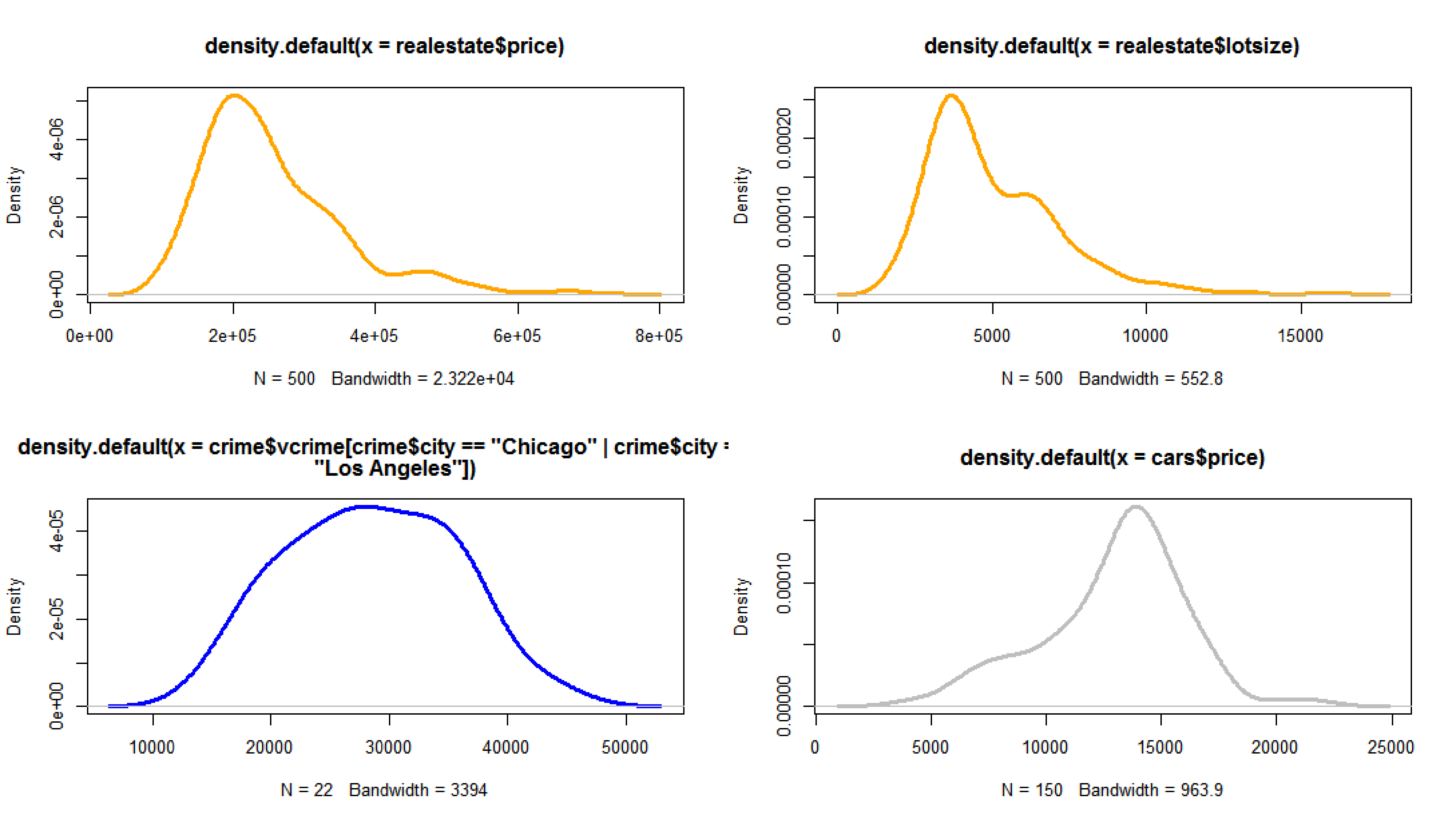
3. Hypothesis testing

4. *Final project discussion



Midterm Grades
mean=85.55
median=90
sd=11.82
Frequency
0
5
10
15
20
25
50
60
70
80
90
100

DISTRIBUTIONS


density.default(x = realestate$price)
Density
0e+00
2e-06
4e-06
0e+00
2e+05
4e+05
6e+05
8e+05
N = 500   Bandwidth = 2.322e+04
density.default(x = realestate$lotsize)
Density
0.00000
0.00010
0.00020
0
5000
10000
15000
N = 500   Bandwidth = 552.8
density.default(x = crime$vcrime[crime$city == "Chicago" | crime$city =
"Los Angeles"])
Density
0e+00
2e-05
4e-05
10000
20000
30000
40000
50000
N = 22   Bandwidth = 3394
density.default(x = cars$price)
Density
0.00000
0.00010
0
5000
10000
15000
20000
25000
N = 150   Bandwidth = 963.9

Distribution functions

Function (prefix)	Description
d	Given a distribution, return the values of Probability Density Function (PDF)
p	Find a probability, given a distribution. This is result from a Cumulative Distribution Function (CDF)
q	Find a quantile, given a distribution. This is a result of the inverse CDF.
r	Given a distribution and relevant parameters (vary by distribution type), randomize a vector of numbers

?Distributions (to obtain information on various distributions found in R)

http://www.statmethods.net/advgraphs/probability.html

OpenIntro Statistics, Chapter 3

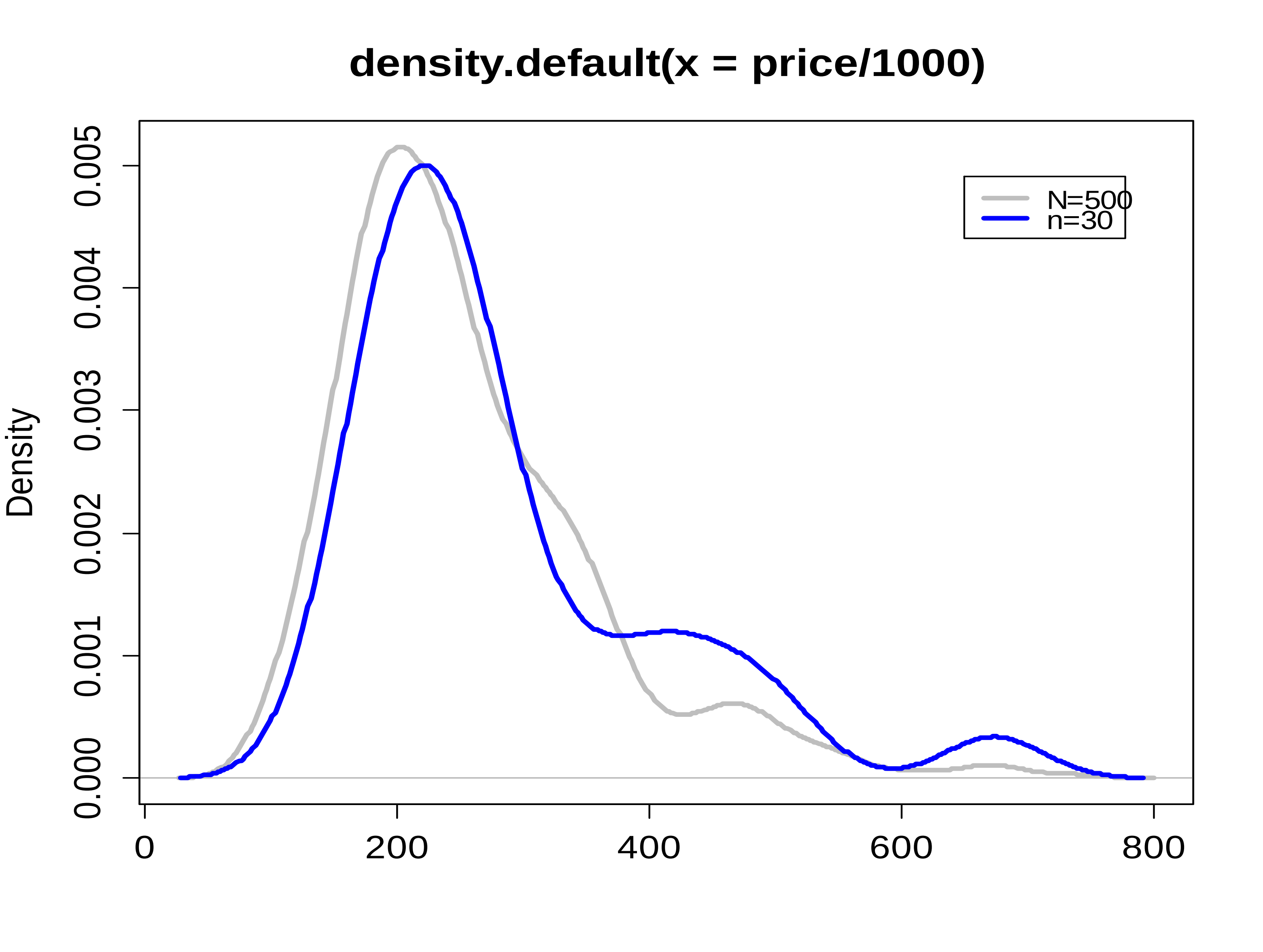
SAMPLING AND CONFIDENCE INTERVALS

Why sampling?

- We typically do not have full data (although this is changing!)
- Instead, we use random samples from the population to estimate parameters.
- We use the following terms to distinguish samples:

  Point estimation = a single parameter estimate (e.g., sample mean)
  Standard error (SE) = sd

Donations example



- Load realestate data
- What is the shape of the price distribution?
- Assign a sample of 30 house prices to **price_samp**. Hint: use set.seed(123). What is the shape of the **price_samp** distribution?
- What is the different between the mean of the entire price column (population mean) and the mean of **price_samp** (sample mean)?
- What is the difference between the following point estimates and the population parameter: median and standard deviation

Sampling assumptions

- ✔ Point estimates from a single random sample, especially a relatively small sample, are often insufficient
- ✔ Estimates taken from many samples will approximate the population parameter.

- Using the previous exercise, increase the sample size to a 100
- By how much are the point estimates more accurate?

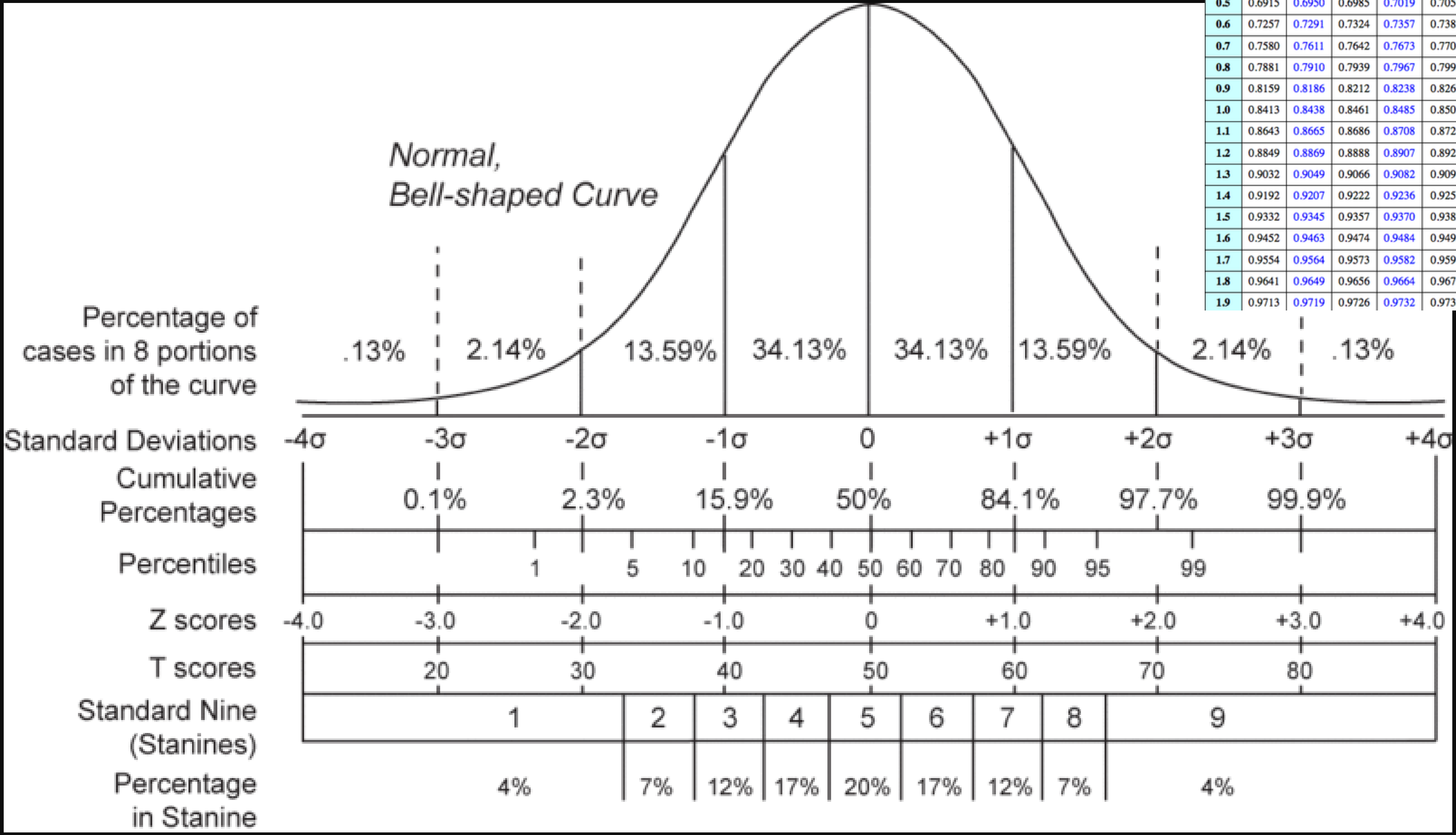
Sampling assumptions (cont.)

- ✔ Estimates taken from <u>many samples</u> will approximate the population parameter.
- ✔ Regardless of the population, estimates will follow a normal distribution.

- Using the previous exercise, continue using n=100, and take 1000 samples (Hint: copy and modify the for loop in the example).
- What is the difference between the means of the sample parameters and the population?
- Are the sample point estimates normally distributed?

Confidence interval

- What it means is confidence level that the actual population parameter is between low-high limits.
- In practical terms: x*SE from point estimate. Common values of x: 95% = $\pm$1.96SE; 99% = $\pm$2.58SE
  (These numbers are known as z-scores).



Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0..6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

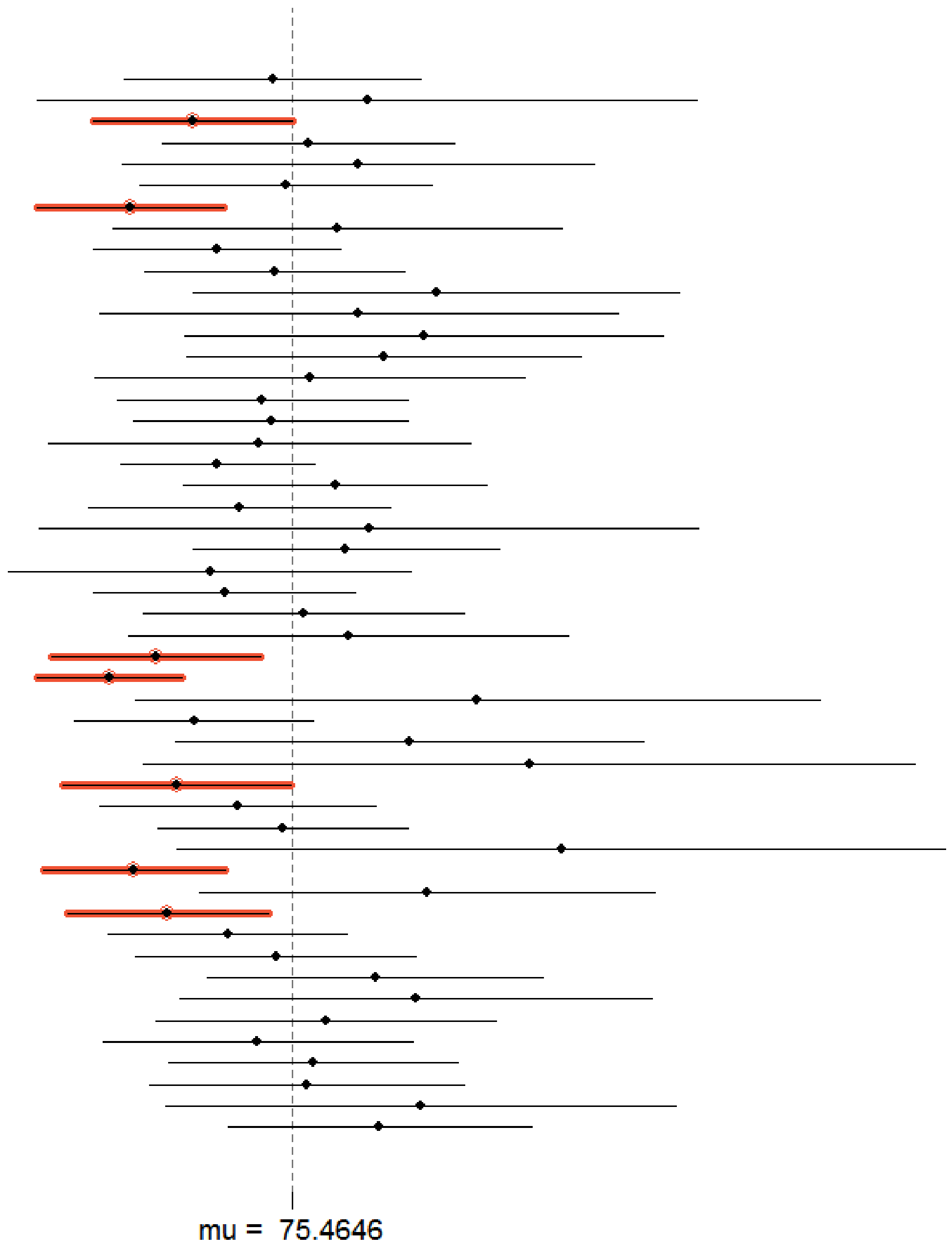
Caveats

- ✔ Sample observations are independent
(Simple random sample with n<10% of N).

- ✔ Sample size is n=>30, but when expecting outliers (as we have seen earlier) n=>100, or more.

- ✔ Population distribution is not very skewed.
(For very skewed distributions, we will need to use methods such as bootstrap. A good example is boot.ci function in the boot package).

Let's try it with a terribly skewed variable, and then examine a "reasonably" skewed variable.


mu = 75.4646

- set.seed(your_uid)
- At a 95% CI, what is the mean home **price** in the **realestate** data?
- At a 99% CI, what is the mean **lotsize** in the realestate data?

HYPOTHESIS TESTING

Why do we use it?

- ✔ Formal test
- ✔ Draws conclusion about the population from a sample

Assumptions

- ✔ DV is measured as an interval (not continuous)
- ✔ Sample is random
- ✔ Observations are independent from one another
- ✔ Population distribution of approximately normal

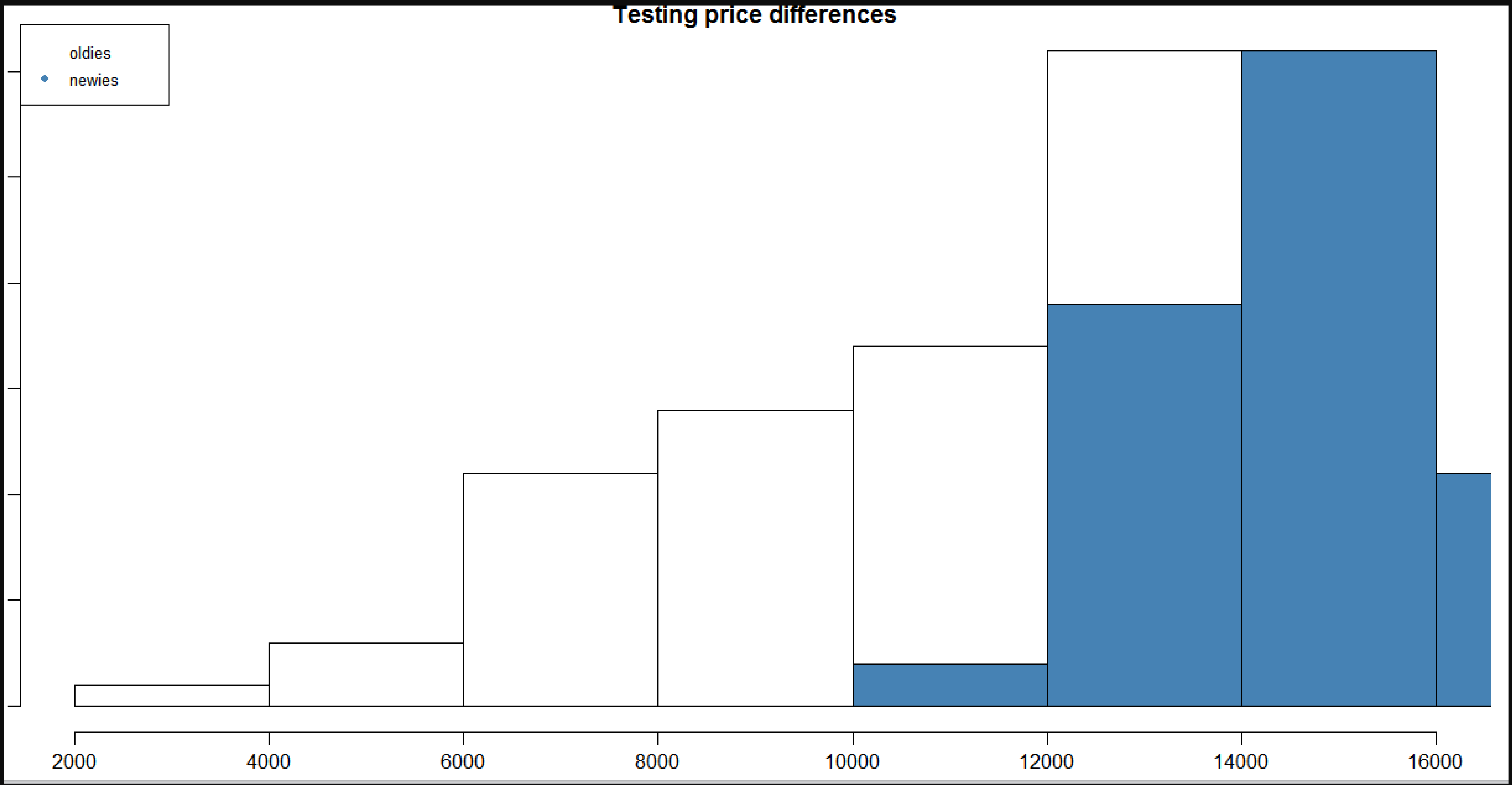
Procedure

- Start by declaring your null hypothesis (H_0) and the alternative hypothesis (H_1)
  (H_1 is your research hypothesis)

- (Opposite of the research hypothesis, e.g., no difference in population mean and a value)

- Determine your α level
  (Typically set at 0.05)

- Interpret the result
  (If $p > \alpha$ cannot reject the null; if $p < \alpha$ reject the null)


Testing price differences
oldies
newies
2000
4000
6000
8000
10000
12000
14000
16000

Do you reject or accept the null hypotheses:

- ✓ **price** of houses with 2 stories is the same as houses with 3 and 4 stories, at α=0.99?
- ✓ **price** of houses with 2 stories is the same as houses with 1 story, at α=0.95?

FINAL PROJECT DISCUSSION

Hypotheses from your projects